A SIMPLE UNIVARIATE OUTLIER IDENTIFICATION PROCEDURE ON RATIO DATA COLLECTED BY THE DEPARTMENT OF REVENUE FOR THE STATE OF KANSAS

by

HYOUNGJIN JUN

B.S., Sungkyunkwan University, 1998

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Approved by:

Major Professor
John E. Boyer, Jr.

Abstract

In order to impose fair taxes on properties, it is required that appraisers annually estimate prices of all the properties in each of the counties in Kansas. The Department of Revenue of Kansas oversees the quality of work of appraisers in each county. The Department of Revenue uses ratio data which is appraisal price divided by sale price for those parcels which are sold during the year as a basis for evaluating the work of the appraisers. They know that there are outliers in these ratio data sets and these outliers can impact their evaluations of the county appraisers.

The Department of Revenue has been using a simple box plot procedure to identify outliers for the previous 10 years. Staff members have questioned whether there might be a need for improvement in the procedure. They considered the possibility of tuning the procedure to depend on distributions and sample sizes. The methodology as a possible solution was suggested by Iglewicz et al. (2007).

In this report, we examine the new methodology and attempt to apply it to ratio data sets provided by the Department of Revenue.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 - Introduction

This project arose out of questions raised by staff members in the Property Valuation Division (PVD) of The Kansas Department of Revenue. That unit is charged with overseeing the work and the reporting of the 105 county assessors in the State of Kansas.

The data sets that the PVD must consider are collections of ratios of assessed values to selling prices of properties which are sold in each county in a given calendar year. It is known that outliers are contained in these data sets and PVD would like to know how best to identify and subsequently remove ("trim") these outliers.

The PVD has a methodology to accomplish this trimming in place and that methodology is automated. In an effort to improve their reporting, the staffs at PVD have sought help and advice as to the quality of their methodology.

This report uses a number of data sets provided by PVD and deemed to be "typical". These are used as the basis for an investigation into the outlier detection methods used by the division.

CHAPTER 2 - Research for Department of Revenue in Kansas

Usage of appraisal prices as basis of taxation

If you live in the United States, you are obligated to pay tax on a variety of bases, whether you want to or not. In particular, more taxes are imposed on expensive items such as real estate. Therefore, taxpayers want the taxes on their properties to be fair and so does the Department of Revenue in Kansas which is in charge of taxation in Kansas.

A primary factor in terms of how much tax should be imposed on properties such as residential and commercial/industrial is the price or value of those properties. Unlike goods in stores, these kinds of properties do not have price tags so determining their values is difficult. Every year appraisers (sometimes they are called assessors) have to consider each property and estimate the value of that property; this process is called mass appraisal.

County assessors are charged with determining the fair market value for all personal property in the state; that is the price the property would bring if it were sold in an open market. This is called the "appraisal" price (value). Then, the market value becomes the basis for personal property taxes that must be paid by property owners. If market value is determined accurately, then the tax burden is shared in a way that is proportional to the owners' assets.

Usage of ratio to study appraised values

One issue consistently arises with appraisal prices. Are those prices really representing true prices? Can we trust their accuracy? Are they evaluated uniformly and fairly? One of the tasks of the Department of Revenue is to evaluate fairness of appraisal prices.

Because each county has its own appraiser who oversees the process in his/her county, the Department of Revenue treats each county as a separate entity. One can think of all appraisal prices in each county as a single population of possible values of all properties. Also we can imagine that there is another method of getting at true values of properties in each county. That would rely on the sales price of those properties which were sold during a given year. Unlike appraisal prices, we cannot acquire all sale prices on all properties in a given year, since not all properties will be sold. In some counties there may be hundreds of properties sold whereas some others might have only a small number per year.

In order to assess the accuracy of the appraised value of personal property, all sales of such property must, by law, be reported to the county assessor of the county in which the property resides. Prices for properties which are sold in a given year then can be used as a measure of the accuracy of the appraisal process.

We have two sets of values to compare and if they turn out to be the same or close to each other, we can conclude that appraisal prices are considered good approximations to the true values (sales price). Therefore, the taxation based on those prices will be conducted on a fair basis.

In studying the performance of the county assessors, the staff at PVD considers ratios of appraisal price to sale price on all properties which are sold in a given year. Values near 1.0 are clearly desirable, since they reflect that the appraisal prices closely approximate the sale prices.

Thus the data sets that the PVD sees are collections of ratios as reported by the county assessors. Some issues that need to be considered in the ratios are discussed in the following sections.

Validity of sale prices as real true values

An assumption that must be made for a ratio to be considered reasonable to include in the study is that the sale price is a fair indication of its value on the open market. Previously, we assumed that sale prices are true values. However we have to be cautious with the definition of true value. There are numerous situations in which we can doubt if the sale price is representing true value of the property.

For example, a wealthy grandfather who loves his grandkids could sell his beautiful mansion whose market value might be one million dollars to his granddaughter for $10,000. Clearly $10,000 is not the true value of the mansion.

Alternatively, if an owner needs cash in a real hurry, he/she would be willing to sell his/her property for much less than the true value.

These kinds of transactions occur and we need to remove them from the database of sale prices which should represent true values.

By contrast, there are some situations that properties are sold for higher prices than their true current market price. If there is strong competition or some information that the value of the

property might be increased in the future due to some forthcoming development, then the sale price would be much higher than its true value at that time.

The determination to classify a sale as valid or invalid is based upon industry standards promulgated by the International Association of Assessing Officers (IAAO). Various factors will be considered during the screening process before an assessor can invalidate a sale. There are a few criteria to check the validity of each transaction so that they can eliminate most obvious cases of invalid sale prices. Some of those criteria are the following and any sale that is affected by the following conditions will be considered invalid for ratio study purposes.

- Property not exposed for sale on the open market
- Transaction was a forced sale
- Sale date not within the current study timeframe
- Uninformed buyer and/or seller
- Sale included an excessive amount of tangible or intangible personal property

Convenience sampling procedure

One notices immediately that considering those properties which are sold during a given year does not generate a random sample, but a convenience sample.

Convenience sampling is not a good way to take a sample but researchers in many fields do it all the time due to limited resources. In our case, properties that are sold comprise the only sample that is available to us. It is completely impossible to make people sell/buy properties randomly. The only option we have with regard to sampling is to use the database of sale prices for each county in a specific year of research.

Ratio analysis as a tool for matched pairs

In order to compare the sales and appraisal prices, it should be noted that one could use matched pairs t-test since each property has two kinds of prices and the goal is to compare the equality of those two values. However, in the field of mass appraisal, ratio analysis is the standard tool because there are large variations between one property and the other. If we use matched pairs t-test, the differences from higher valued properties might well override the rest and the conclusion of overall accuracy of appraisal price will be biased.

An Assessed Value-to-Sales Price Ratio study is the international standard for determining the accuracy of a mass appraisal. The definition of the ratio is the following:

$$Ratio = \frac{appraisal\ price}{sale\ price} \quad (1)$$

Statistics for ratio analysis: COD and PRD

In examining the samples of ratios, the Department of Revenue computes some summary statistics on those ratios. Coefficient of Dispersion (COD) and Price-Related Differential (PRD) are two primary ones.

Coefficient of Dispersion (COD)

$$COD = \frac{Average\ Absolute\ Deviation\ from\ Median\ ratio}{Median\ ratio} \times 100$$

COD is the most common measure of uniformity in the mass appraisal industry. It is similar to coefficient of variation (CV). COD measures the average amount of dispersion from the median ratio and expresses it as a percentage of the median ratio. The statistic indicates how closely the ratios are clustered around the median ratio.

The lower the coefficient of dispersion, the more uniform the assessments. A high COD suggest a lack of uniformity. The ideal value is 0, however, this cannot be considered a realistic goal in an imperfect real estate market.

Price-Related Differential (PRD)

$$\boldsymbol{PRD = \frac{Arithmetic\ mean\ ratio}{Weighted\ mean\ ratio}}$$

Weighted mean ratio is obtained by dividing the total sum of appraisal prices by the total sum of sale prices.

The price-related differential (PRD) is a statistic for measuring assessment regressivity or progressivity. Appraisals are considered regressive if high-value properties are underappraised relative to low-value properties and progressive if high-value properties are relatively overappraised.

A PRD of 1.0 is the most desirable state and indicates that no assessment bias exists between the low and high value properties. A PRD greater than 1.0 suggests that high-value property may be underappraised relative to lower valued property. If the PRD is less than 1.0, high-value property may be overappraised relative to low-valued property. The International Association of Assessing Officers Standard on Ratio Studies recommends the PRD should range between 0.98 and 1.03.

Significance of outlier identification

In general, screening data for outliers is an important, but often underutilized, part of a careful statistical investigation. Iglewicz and Hoaglin (2007) state that, "We recommend that data be routinely inspected for outliers, because outliers can provide useful information about the data."

Outliers may represent a different population than the original population, so it is important to remove them before we conduct any statistical analyses on data to reach any conclusion about the original data sets.

Even though we have a few criteria to identify invalid sale prices, there is still the possibility of a sale being invalid for reasons unknown to us. This is clear, since occasionally ratios (defined by (1)) as large as 900% are found in the studies. Therefore, it is a very important to identify outliers among those ratios to investigate the overall accuracy of appraisal prices. The outlier identification procedure is considered to be an important matter in the analysis of the validity of sales prices; it allows us to identify problems and abnormalities that the sales criteria did not identify.

CHAPTER 3 - Current Outlier identification procedure of the Department of Revenue in Kansas and used data sets

Outlier identification procedure of PVD

In the 1950's the ratio study was conducted on only the data found between the first and third quartiles in each county's data. This means that 50% of data was thrown away. By the 1980's a 5% trim on "large" samples was suggested as a part of outlier trimming guidelines proposed by the IAAO (International Association of Assessing Officers). In 1992 several modifications were made in the PVD sales ratio study. This included a formula driven outlier identification/trimming process. This method was incorporated in the IAAO Standard on Ratio Studies in 1999.

The methodology adopted by PVD was an outlier identification procedure based on boxplots. This procedure uses the interquartile range and some arbitrary factors to determine cutoff values for outliers. In order to find the interquartile range, we need to find the first and third quartiles. There are several ways to find quartiles and the following is an example of the trimming procedure adopted by PVD to cut off outlying ratios.

Suppose, for example, a data set of 10 ratios contains the following values.

45.7 63.4 71.7 77.6 81.0 83.3 91.6 103.7 115.5 171.9

1. First sort the array of ratios from low to high.

2. To locate the first and third quartile points.
   Formula used by PVD
   The first quartile point is: (n * .25) + .25, (where n is the sample size).
   The third quartile point is: (n * .75) + .75.

3. For this example, n=10. Therefore,

First quartile: (10 * .25) + .25 = 2.75

The first quartile is located between observation 2 and 3

4. Interpolation:

   71.7 - 63.4 = 8.3 * .75 = 6.2

   63.4 + 6.2 = **69.6**

5. Third quartile: (10 * .75) + .75 = 8.25

   Third quartile is located between observation 8 and 9

6. Interpolation:

   115.5 – 103.7 = 11.8 * .25 = 3.0

   103.7 + 3.0 = **106.7**

7. Therefore the interquartile range, i.e., distance between the third quartile and the first quartile, is to deemed to be 37.1.

8. Next, the interquartile range is multiplied by a factor of 1.5 to establish a base width for general outlier trimming.

   To find the trim width, we multiply 37.1 * 1.5 = 55.7

9. The last step is to subtract the trim width from the first quartile and add it to the third quartile. This establishes where the trim points are located.

   Lower trim point: 69.6 – 55.7 = 13.9

   Upper trim point: 106.7 + 55.7 = 162.4

Any ratios below **13.9** or above **162.4** would be termed outliers and would be "trimmed" or removed from further consideration.

Notice that this is a simple outlier identification procedure based on a boxplot except that PVD added 0.25 and 0.75 to calculate the first and third quartiles, respectively. PVD applied this

modification to a common boxplot methodology to compensate for the "discreteness" of ratio data that PVD deals with.

Questions from PVD regarding the Boxplot methodology

PVD believes that contaminated data points are always possible in the ratio data. They may result for a few possible reasons, including a sale price misstatement on the Sales Validation Questionnaire, missing critical validation or adjustment factors on the questionnaire, a clerical error resulting in an incorrect parcel reference or classification, or a data entry error.

PVD has been using the Boxplot trimming procedure with the factor of 1.5 and a maximum trim of 20% since 1999. By doing so, PVD was hoping to catch most of contaminated data points from sample ratio data. However, PVD has been questioning the validity of these two relatively arbitrary factor values and considering possible improvement on the methodology. Actually the IAAO recommendations suggest a maximum trim of five to ten percent for larger sample sizes.

The outlier trimming procedure is important because of the strong dependency of the coefficient of dispersion (COD) and price related differential (PRD) on extreme ratios in the samples. Acting conservatively when eliminating outliers may cause poor scores on these two statistics. Consequently PVD may judge an appraiser's ability negatively in a particular county, even if the appraiser has done an adequate or good job.

The other extreme case is when we remove too many data points by setting those two factors higher than we should. In this case, PVD obtains scores on COD that are too good to be true based on experience. Therefore COD is a good indicator to decide the threshold for two factors and it has been an important issue within IAAO as to how low the COD can be. The general consensus is about 5%, but nobody knows for sure and unfortunately there is little empirical data available.

The following summarizes questions brought up by PVD for the Technical Advisory Committee regarding outlier trimming procedures,

- Is the number of counties with a low COD a concern?
    - A low COD can result from excessive trimming of outliers.

- Is the outlier trimming process too generous?
- Is there any way to establish a market noise threshold?
- Should some factor such as a significance level of α be implemented to adjust a trimming rate?
- Should the maximum trim percentage simply be lowered?
- Should the maximum trim percentage be tempered by sample size?

A suggested improved boxplot methodology

In effort to answer some of questions brought up by PVD, the modified boxplot outlier identification procedure by Iglewicz and Banerjee(2007) was suggested.

Both this procedure and the current method used by PVD are based on using boxplots to identify outliers. In outlier identification procedures using boxplots, we need to find both upper limits and lower limits to claim outliers. They use IQR (interquartile range) times a multiplying factor such as 1.5 from upper quartiles as their upper limits. Their lower limits are determined by IQR times a multiplying factor from lower quartiles. The critical difference between this new procedure and the current one used by PVD is how to decide multiplying factors such as 1.5. The current boxplot procedure of PVD uses the factor of 1.5 regardless of sample sizes and population distributions, whereas Iglewicz and Banerjee(2007) suggest that the factors depend on both sample size and distribution.

The idea was simple. They wanted to fix the probability of claiming outliers out of random sample from known distributions at a specific value, say, $\alpha=0.05$. It was found that the probability of identifying false outliers increases as sample sizes increase with the original boxplot procedure. In order to keep this probability constant, we need to have different values for these multiplying factors depending on sample sizes. Consequently, we will have certain level of confidence that we are fairly performing outlier trimming procedures on different sized samples.

The following is the brief summary overview of this procedure. The method is based on the boxplot procedure which allows us to identify outliers if data point falls outside the interval $(Q_1-g*(Q_3-Q_1), Q_3+g*(Q_3-Q_1))$. They use this same formula for symmetric distributions and for

skewed ones they modified the interval as the following. (Q_1-g*(M-Q_1), Q_3+g*(Q_3-M)). Note that Q_1 is the lower quartile, Q_3 is the upper quartile, M is the median and g is a multiplying factor. We can calculate Q_1, Q_3 and M from the random sample we wish to trim.

Finding g is straightforward once we establish the equation for the probability of identifying outliers at α level. The preceding sentence can then be written as the following equations. For symmetric distributions, $P(X_{(n)} > Q_3+g*(Q_3-Q_1))=\alpha/2$ and for skewed ones $P(X_{(n)} > Q_3+g*(Q_3-M))=\alpha$ as long as our concern is only upper outliers. Assuming that we know the underlying distributions and that the sample size is large enough, these equations can be solved for g as follows by replacing quartiles with the corresponding inverse cumulative functions.

For symmetric distributions,

$$g = \frac{(F^{-1}\left((1-\frac{\alpha}{2})^{\frac{1}{n}}\right) - F^{-1}(.75))}{(F^{-1}(.75) - F^{-1}(.25))}$$

For skewed distributions,

$$g = \frac{(F^{-1}\left((1-\alpha)^{\frac{1}{n}}\right) - F^{-1}(.75))}{(F^{-1}(.75) - F^{-1}(.5))}$$

The above equations are only valid when sample sizes are large enough. Iglewicz et al. (2007) found that when sample sizes are approximately 2000 or more, we can use those equations. Additionally, they calculated a modifying factor k for smaller sample sizes for well-known distributions such as normal, Gamma, t, and Weibull. So instead of using g, we use kg. Therefore, both g and k are functions of sample size and depend on the underlying distribution.

In order to see how well this new method would work for our purpose, first we need to determine if ratio data sets from PVD are distributed similarly to one of distributions such as normal, gamma, t and Weibull, which is one of primary goals of this report.

Data sets used in this report

Staff from the Department of Revenue provided 16 data sets which came from different counties in Kansas. An experimental unit will be an individual sale in the county in the given year. Sample sizes range from 17 to 394. Each data set has 15 variables including sale date, sale price, appraised price, and ratio.

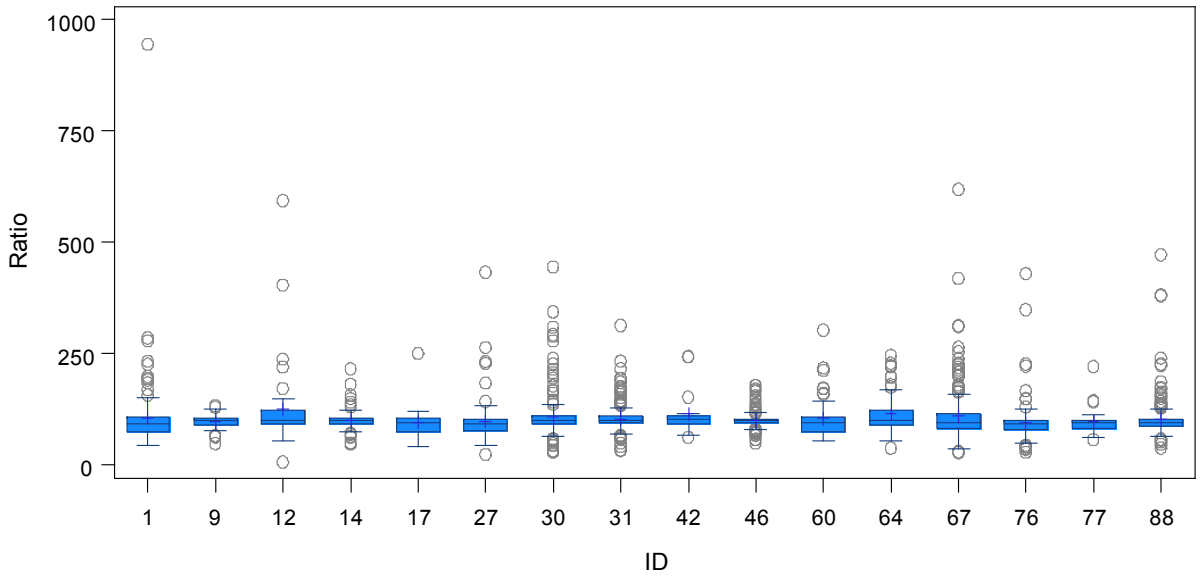
CHAPTER 4 - Efforts

Attempts to fit normal, Gamma, and Weibull distributions

Boxplots for overview of whole ratio data sets

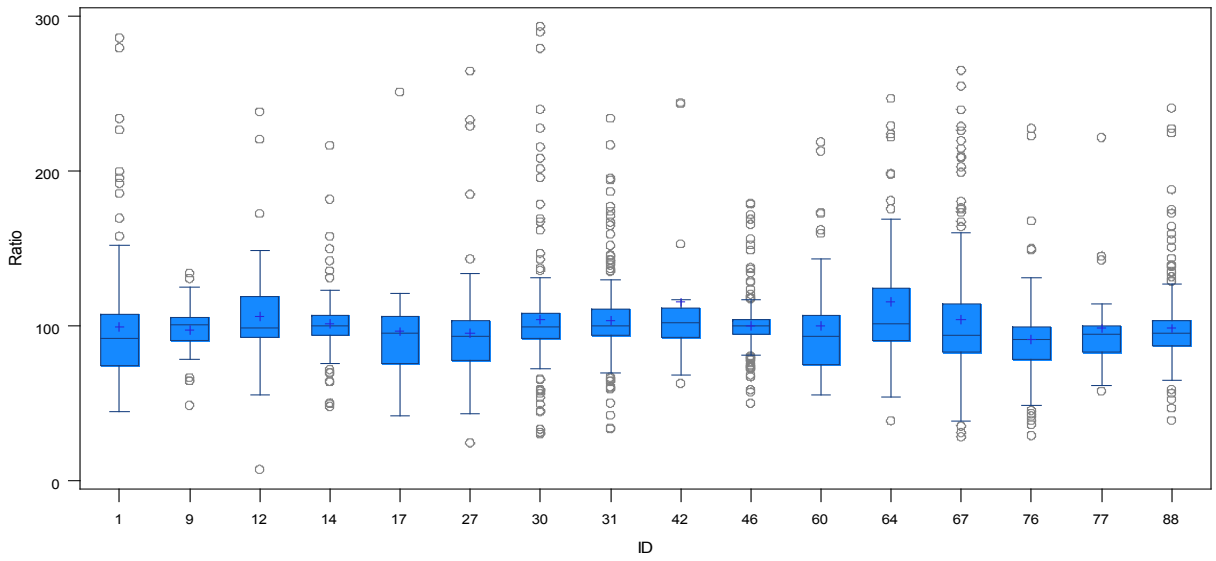
Box plots are not only used for outlier identification tools but also convenient tools for comparing distributions of a quantitative variable across levels of a grouping variable. The boxplot charts below are derived from 16 sets of ratio data from 16 counties in Kansas.

Figure 1: Boxplot of 16 county ratio data



The extreme values in these data sets make it so that we cannot see the distribution characteristics of the main parts of the data. In order to see a little clearer image of each distribution, a maximum scale is cut down to 300 in Figure 2.

Figure 2: Boxplot of 16 county ratio data (<300)



From these graphs, we can notice a few of aspects of distributions. They tend to be skewed to the right and condensed around ratio=100. There are outliers on both sides of 100 but there tend to be more large outliers than small.

In order to use the modified boxplot outlier identification procedure, it would be useful to be able to identify distributions as one of several well known distributions such as normal, Gamma, t or Weibull. The boxplots show us that these distributions tend to be skewed to the right, which suggests that normal distribution might not be the best choice. However, it will be worthwhile if we can find any connection with normal distribution.

Attempt to fit normal and t-distribution to data sets

The following (Figure 3 and Figure 4) are two typical histograms from 16 data sets. The histograms themselves, with the fitted normal curves superimposed, show us that normal distributions are not good fits. As we would expect from a visual examination of the data, all the goodness-of-fit tests for normality such as Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling and Cramer-von Mises failed for all 16 data sets.

Figure 5 and 6 are the QQ plots for these two data sets assuming normality of underlying populations. Again we can see normality does not seem to be a reasonable model for these data. Notice that slopes of both curves are increasing from left to right which means that they are skewed to the right.

From the histograms with fit to normal curves, we notice that ratio data sets are condensed around ratio=100 instead of having thick tails. The t-distribution has thicker tails than normal distribution and fitting t-distribution would work better if the data sets have thick tails. Therefore, we can deduce that fitting a t-distribution would not work based on the facts that normal fitting does not work and ratio data sets appear to have thin tails.

Figure 3: Histogram 1 with normal fitting

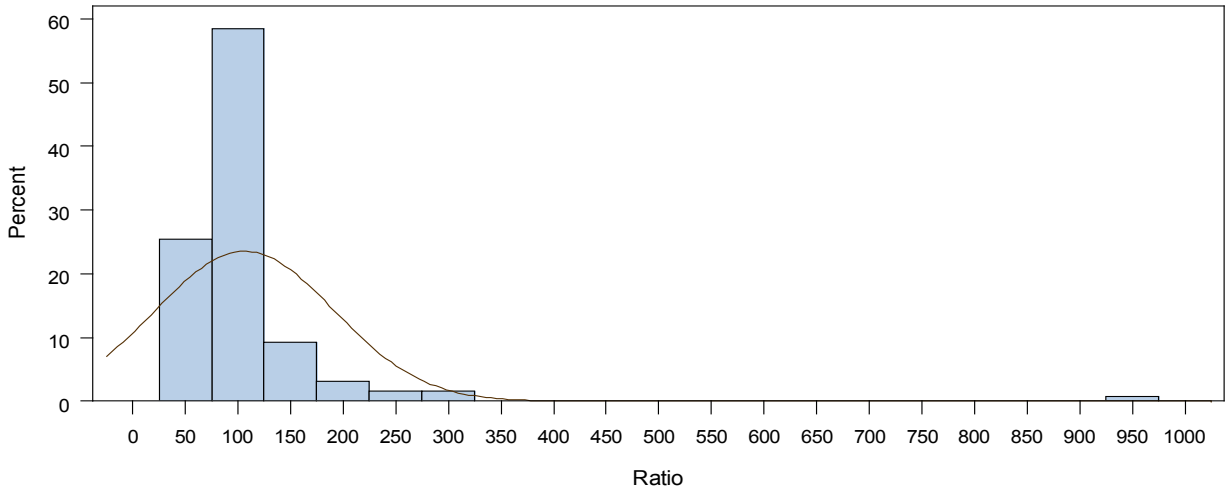


Figure 4: Histogram 2 with normal fitting

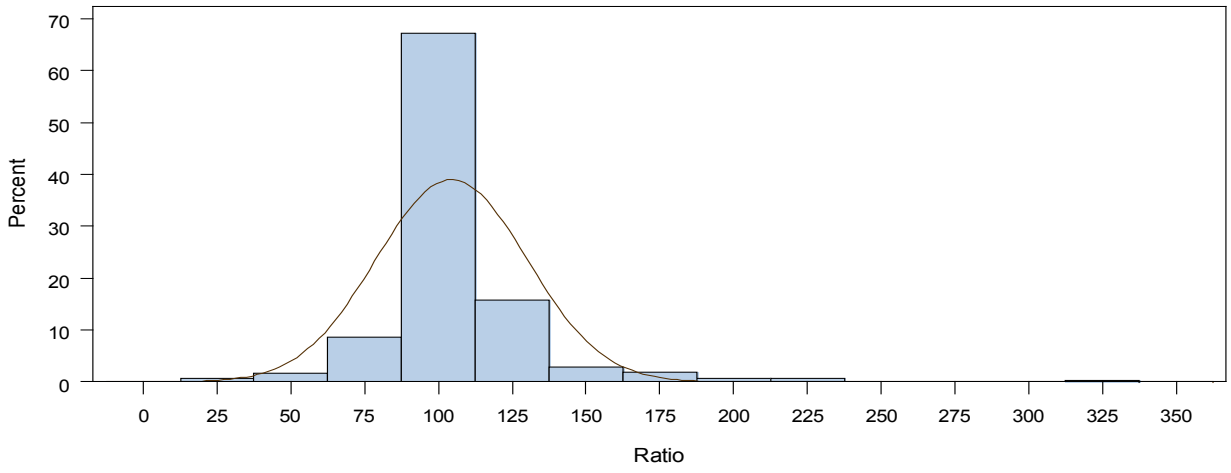


Figure 5: Q-Q Plot 1

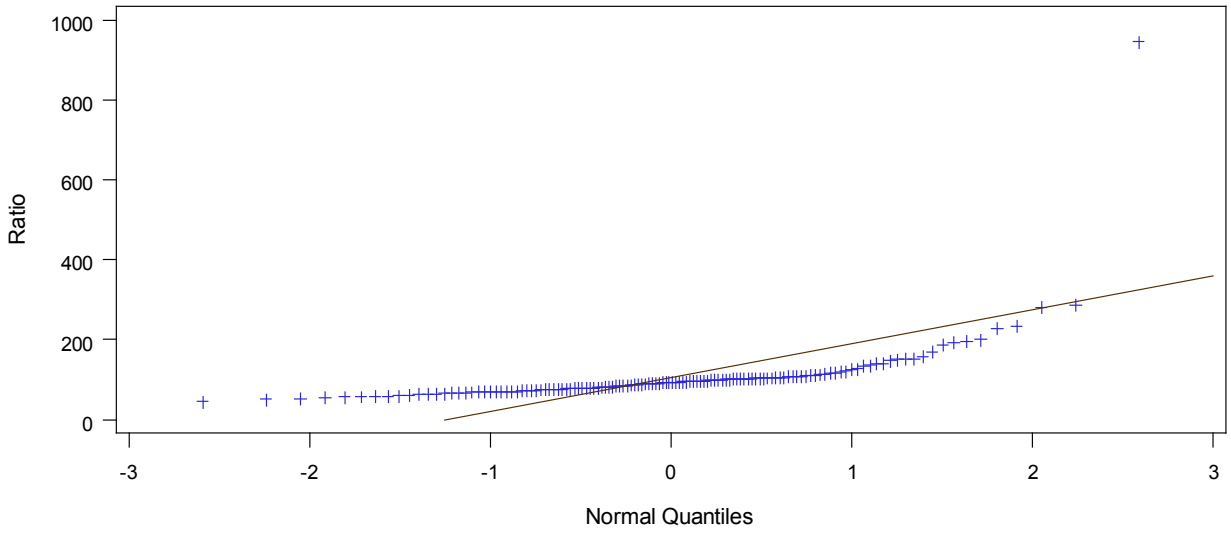
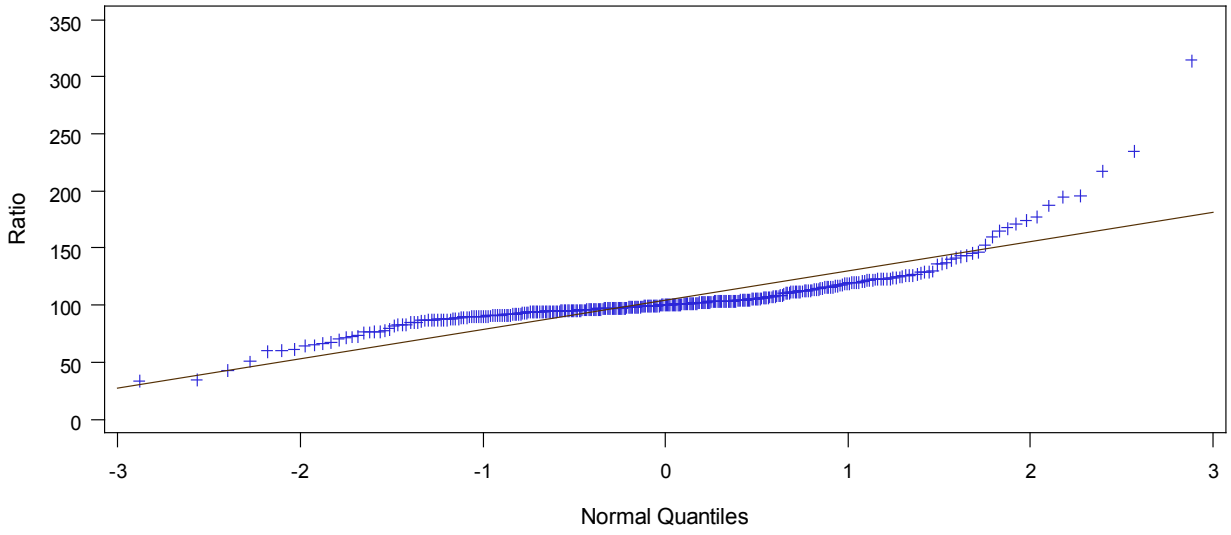


Figure 6: Q-Q Plot 2



Attempt to fit Gamma, Weibull

Next we attempted to fit Gamma and Weibull distribution to the data sets. All goodness-of-fit tests failed at 5% level of significance. The following figures are fitted Gamma and Weibull distributions on two typical histograms of the ratio data sets. It is obvious that they do not fit well.

Figure 7: Histogram 3 with Weibull and Gamma fitting

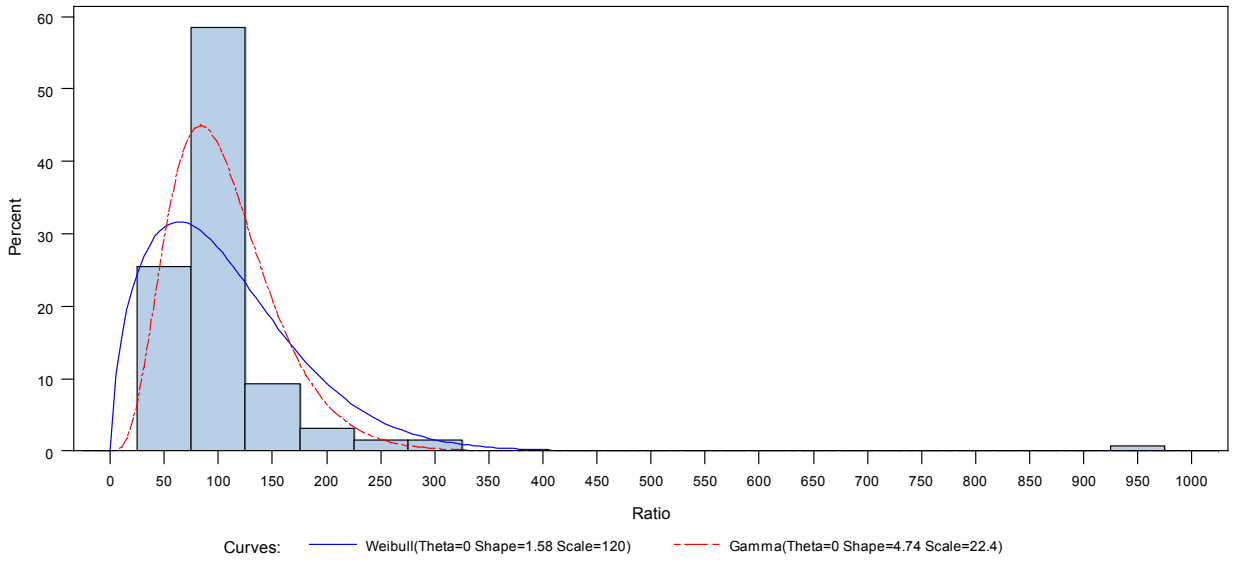
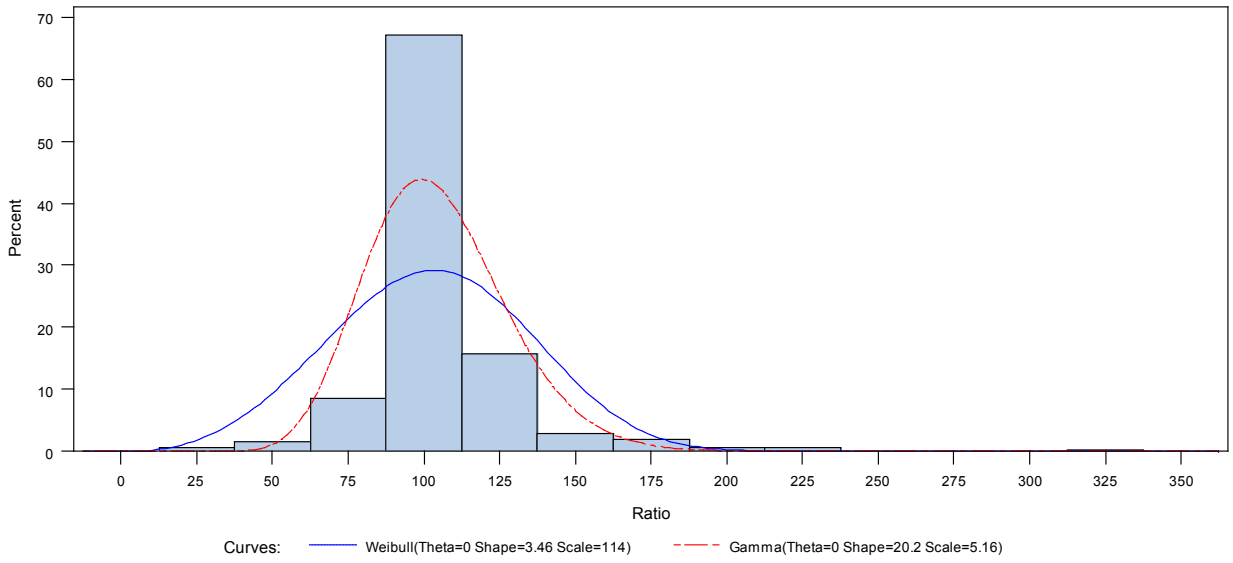


Figure 8: Histogram 4 with Weibull and Gamma fitting



Attempts to trim outliers using g depending on distributions such as normal, Gamma, Weibull and t distributions

Once it was deduced that the data sets did not seem to fit normal, t, Weibull and Gamma distributions for the given ratio data sets, we attempted to determine what percentages of outliers are trimmed using g's from different distributions on those data sets. In addition, we sought to determine which distribution assumption gives the most consistent trimming rate over 16 data sets. We would hope that the best candidate would have a consistent 5%-10% trimming rate with neither too high nor too low a trimming rate.

Need for modified formulas for g's which do not depend on skewness

The formulas Iglewicz et al used to obtain g's are the following:

For symmetric distributions,

$$g = \frac{(F^{-1}\left((1-\frac{\alpha}{2})^{\frac{1}{n}}\right) - F^{-1}(.75))}{(F^{-1}(.75) - F^{-1}(.25))}$$

For skewed distributions,

$$g = \frac{(F^{-1}\left((1-\alpha)^{\frac{1}{n}}\right) - F^{-1}(.75))}{(F^{-1}(.75) - F^{-1}(.5))}$$

In order to use these formulas, we need to know if a fitted distribution is either symmetric or skewed. In addition to this, the formula for skewed distributions assumes only upper outliers. However, PVD believes that there may be outliers in both tails.

A small modification is made to solve the above problems. Basically, we used the formula for skewed distributions. To consider both tail outliers, we had to calculate g_{up} and g_{lo}. To get fixed α, α/2 is applied to both g_{up} and g_{lo}. g_{up} is obtained by

$$g_{up} = \frac{(F^{-1}\left((1-\frac{\alpha}{2})^{\frac{1}{n}}\right) - F^{-1}(.75))}{(F^{-1}(.75) - F^{-1}(.5))}$$

The following is how to obtain g_{lo}.

$$P(X_{(1)} < F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right)) = \frac{\alpha}{2}$$

$$P\left(X_{(1)} < F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right)\right)$$

$$= 1 - P(X_{(1)} \geq F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right))$$

$$= 1 - [1 - F(F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right))]^n$$

$$= \frac{\alpha}{2}$$

Therefore,

$$[1 - F(F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right))]^n = 1 - \frac{\alpha}{2}$$

$$1 - F\left(F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right)\right) = (1 - \frac{\alpha}{2})^{\frac{1}{n}}$$

$$F\left(F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right)\right) = 1 - (1 - \frac{\alpha}{2})^{\frac{1}{n}}$$

$$F^{-1}(0.25) - g_{lo} * \left(F^{-1}(0.5) - F^{-1}(0.25)\right) = F^{-1}[1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{n}}]$$

Solving for g_{lo}, we get the following.

$$g_{lo} = \frac{F^{-1}(0.25) - F^{-1}[1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{n}}]}{F^{-1}(0.5) - F^{-1}(0.25)}$$

Note that for our study, we fix $\alpha = 0.05$.

Actual values of g_{up} and g_{lo} for normal, Gamma, Weibull and t distributions

By applying actual values of g_{up} and g_{lo} to (Q_1- g_{lo}*(M-Q_1), Q_3+ g_{up}*(Q_3-M)), we can find upper limits and lower limits for outlier identification.

We chose a couple of shape parameters for Gamma and Weibull distributions and two degrees of freedom for t distribution for our study. The following table summarizes values of g_{up} and g_{lo} by those distributions with parameters and by sample sizes of 100, 200 and 300. The last row shows averaged values of g_{up} and g_{lo} over those three sample sizes.

Table 1: values of g_{up} and g_{lo}

	normal		**gamma**						**Weibull**				**t**			
			shape=1		*shape=3*		*shape=5*		*shape=2*		*shape=4*		*df=4*		*df=10*	
	g_{up}	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**	**g_{up}**	**g_{lo}**
n=100	4.2	4.2	9.9	0.7	7.2	1.7	6.4	2.2	4.9	1.8	3.5	3.4	12.9	12.9	6.2	6.2
n=200	4.4	4.4	10.9	0.7	7.8	1.7	7.0	2.2	5.3	1.8	3.7	3.5	15.6	15.6	6.9	6.9
n=300	4.6	4.6	11.5	0.7	8.2	1.7	7.3	2.3	5.5	1.8	3.9	3.5	17.4	17.4	7.3	7.3
average	4.4	4.4	10.8	0.7	7.7	1.7	6.9	2.2	5.2	1.8	3.7	3.5	15.3	15.3	6.8	6.8

By using those averaged values of g_{up} and g_{lo}, Q1(the first quartile), M(median) and Q3(the third quartile) for each ratio data set, we can calculate UL(upper limit) and LL(lower limit) to trim outliers in both tails. Then, we counted the number of outliers in each tails, added them up to get total numbers and calculated percentages of total outliers in data sets by dividing total numbers of outliers by sample sizes. The following shows the result of those computations for the normal distribution.

Table 2: Percentages of outliers in 16 ratio data sets assuming normal distribution

							Numbers of Outliers			
File ID	N	Q1	M	Q3	UL	LL	N upper	N lower	N total	%
1	130	74.7	92.3	107.6	174.5	-2.7	9	0	9	**6.9**
9	34	90.7	100.8	105.9	128.3	46.5	2	0	2	**5.9**
12	42	93.0	100.0	124.0	229.1	62.4	3	3	6	**14.3**
14	81	94.0	100.0	107.0	137.5	67.6	5	5	10	**12.3**
17	25	76.2	95.3	106.6	156.0	-7.5	1	0	1	**4.0**
27	96	78.9	93.3	104.0	150.9	15.7	5	0	5	**5.2**
30	236	92.1	99.6	110.1	156.3	59.0	16	10	26	**11.0**
31	316	94.5	100.2	110.9	157.7	69.3	12	11	23	**7.3**
42	17	93.0	102.5	112.1	154.0	51.3	2	0	2	**11.8**
46	394	94.9	100.0	104.1	121.8	72.6	14	6	20	**5.1**
60	48	75.2	94.5	107.4	163.7	-9.7	5	0	5	**10.4**
64	60	91.1	102.0	124.8	224.9	43.7	2	1	3	**5.0**
67	186	83.9	94.4	116.1	211.3	37.7	11	3	14	**7.5**
76	149	79.5	91.8	100.0	135.6	25.6	7	0	7	**4.7**
77	24	83.6	94.5	100.0	124.0	35.8	3	0	3	**12.5**
88	265	87.7	95.6	104.2	141.7	53.1	14	3	17	**6.4**

In the above table only the assumption of normal distribution is used to obtain cutoff values for 16 ratio data sets. By using similar procedures assuming other distribution, we calculated all the percentages of outliers in data sets for each distribution assumption. Table 3 summarizes those results.

We have argued that the best choice will be a procedure which produces a consistent 5%-10% trimming rate. From Table 3, Gamma distributions with all three shapes give some percentages well over 20%. The Weibull distribution with shape parameter of 2 is similar to Gamma. However, Weibull distribution with shape parameter of 4 seems nearly as good as normal distribution. The t distribution with degree of freedom of 10 seems good in terms of average value, but there is a zero trimming rate on a couple of data sets. Overall, the normal distribution assumption seems the best. In order to support that conclusion, means and standard deviations of the trimmed percentages for each of the 16 data sets for each distribution assumption are provided in Table 4.

Table 3: Percentages of outliers in 16 ratio data sets assuming normal, Gamma, Weibull and t distributions

		normal	Gamma			Weibull		t	
			shape=1	shape=3	shape=5	shape=2	shape=4	df=4	df=10
File ID	N	%	%	%	%	%	%	%	%
1	130	6.9	10.0	3.8	3.8	6.2	7.7	0.8	3.8
9	34	5.9	14.7	11.8	11.8	14.7	11.8	0.0	0.0
12	42	14.3	19.0	16.7	16.7	16.7	16.7	2.4	7.1
14	81	12.3	19.8	17.3	16.0	19.8	17.3	1.2	7.4
17	25	4.0	8.0	8.0	4.0	8.0	4.0	0.0	4.0
27	96	5.2	15.6	10.4	9.4	11.5	7.3	1.0	5.2
30	236	11.0	19.9	16.1	12.7	17.4	11.9	2.5	6.4
31	316	7.3	16.8	9.5	8.5	11.1	8.9	0.3	3.2
42	17	11.8	35.3	29.4	29.4	29.4	17.6	0.0	11.8
46	394	5.1	17.3	10.2	9.4	10.9	6.9	1.0	3.6
60	48	10.4	10.4	6.3	6.3	6.3	14.6	2.1	6.3
64	60	5.0	18.3	6.7	5.0	8.3	8.3	0.0	0.0
67	186	7.5	16.1	11.8	6.5	12.4	11.8	0.5	2.7
76	149	4.7	15.4	9.4	8.1	10.7	6.7	2.0	3.4
77	24	12.5	25.0	20.8	16.7	20.8	12.5	4.2	12.5
88	265	6.4	16.6	10.2	7.9	10.9	8.3	1.5	3.8

Table 4: Means and Standard Deviations of trimming rates over 16 data sets assuming normal, Gamma, Weibull and t distribution

	normal	Gamma			Weibull		t	
		shape=1	shape=3	shape=5	shape=2	shape=4	df=4	df=10
Mean	8.15	17.39	12.39	10.76	13.44	10.76	1.23	5.06
Stdev	3.35	6.34	6.39	6.53	6.15	4.17	1.18	3.49

From Table 4, only the normal and t distribution with df=10 meet the criterion of having a trimming rate being between 5% and 10%. We found out that Weibull with shape=4 is not quite as good as the normal assumption because the average trimming rate is greater than that for normal distribution and so is standard deviation . Again, t distributions have multiple zero trimming rates. Therefore, we concluded that the normal assumption is the best choice for distribution assumption to obtain g_{lo} and g_{up}.

After we concluded that normal assumption is the best, we wanted to see whether it is better to use average g or g which depends on sample sizes. Instead of taking an individual sample size into consideration to obtain g_{up} and g_{lo}, we chose representative sample sizes. For example, sample size of 100 is chosen for data sets of sizes less than 150. Sample size of 200 is chosen for data sets of sizes between 150 and 250. Finally, 300 is used for data sets of sizes greater or equal to 250. Table 5 shows trimming rates for 16 ratio data sets using both methods.

Table 5: Percentages of outliers in 16 data sets assuming normal distribution by average g and g dependent on N

File ID	N	average g %		g dependent on N %
42	17	11.8		17.6
77	24	12.5		12.5
17	25	4.0		4.0
9	34	5.9		5.9
12	42	14.3	N <150	14.3
60	48	10.4	$g_{up}=g_{lo}=4.2$	12.5
64	60	5.0		8.3
14	81	12.3		13.6
27	96	5.2		5.2
1	130	6.9		6.9
76	149	4.7		4.7
67	186	7.5	150≤N <250	7.5
30	236	11.0	$g_{up}=g_{lo}=4.4$	10.6
88	265	6.4	N≥250	6.0
31	316	7.3	$g_{up}=g_{lo}=4.6$	7.0
46	394	5.1		4.8

In Table 5, notice that trimming rates of "average g" for smaller data sets tend to be smaller than ones of "g dependent on N". For bigger data sets, it's opposite. This is due to the fact that "g dependent on N" is smaller than "average g" for small sample sizes, which results in identifying more outliers. The similar explanation can be used for bigger sample sizes.

For bigger sample sizes, trimming rates by both "average g" and "g dependent on N" are not much different. However, notice that for smaller sample sizes trimming rates using "g

dependent on N" generated high trimming rates for our standard. For this reason, we advise the staff in PVD to use "average g" instead of "g dependent on N".

CHAPTER 5 - Conclusions

Based on goodness-of-fit tests as well as visual and logical examination of histograms with fitted distributions, we concluded that normal, t, Gamma or Weibull do not fit ratio data sets provided by the Department of Revenue particularly well.

Instead of discarding this suggested outlier identification procedure, we proceeded to find out what distribution assumption among normal, t, Gamma and Weibull is the best candidate based on the assumption that the trimming procedure using the best one should maintain a consistent 5%-10% trimming rate with neither too high nor too low a trimming rate.

Our study found the normal distribution to be the best assumption, providing an average trimming rate of 8.2% with standard deviation of 3.4 for these 16 ratio data sets. Some of distributions gave too high trimming rates or too low and others were less consistent than normal distribution.

The staff in the Department of Revenue has been using g=1.5 for the boxplot outlier identification procedure. They have not used any adjustment due to sample size or distribution of data sets. In our study, g of 1.5 seemed to be too small, possibly trimming too much, thus resulting in a COD that is too good to be true. It definitely generates larger trim rates than 8.2%. Notice that in our study, the factor g was found to be 2.2 using normal distribution and it gave an average trim rate of 8.2%.

From our study, we conclude that the current outlier trimming procedure of the Department of Revenue in Kansas may be more extreme than it needs to be. In our opinion the factor g that the staff in PVD is using should be reconsidered and we suggest that it should be increased to around 2 from 1.5, resulting in a lower trimming proportion.

The subject could be researched further in the future. We offer the following suggestions.

The first one is that a reasonable trimming rate for ratio data sets should be determined on more scientific basis. In addition to it, adequate ranges of two main statistics of ratio data, COD (Coefficient of Dispersion) and PRD (price-related differential) should be obtained. The trimming rate affects the values of both COD and PRD. They are kind of two sides of a coin.

The next suggestion is that we continue the search for the appropriate underlying distribution. Since all ratio data sets seem to be skewed to the right, we could try to fit them into a lognormal distribution and then transform them into a normal distribution. At that point, we may be able to apply the simple univariate outlier identification procedure using the normal distribution.

References

2009 Kansas Real Estate Ratio Study

Iglewicz, Boris and Banerjee, Sharmila (2001). "A Simple Univariate Outlier Identification Procedure", Proceedings of the Annual Meeting of the American Statistical Association.

Iglewicz, Boris and Banerjee, Sharmila (2007). "A Simple Univariate Outlier Identification Procedure Designed for Large Samples", Communications in Statistics-Simulation and Computation, 36:2, 249-263.

Hoaglin, D.C., and Iglewicz, B. (1987), "Fine Tuning Some Resistant Rules for Outlier Labeling", Journal of American Statistical Association. 82, 1147-1149.

Hoagline, D.C., Iglewicz, B., and Tukey, J.W. (1986), "Performance of Some Resistant Rules for Outlier Labeling", Journal of American Statistical Association. 81, 991-999.

Tukey, J.W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.